The Geography of Twitter Topics relating to the UK EU Referendum 2016

Robert Webster

January 11, 2018

Summary
This paper uses topic modelling of tweets relating to the UK EU Referendum from the campaign period in 2016. Tweets were classified into topics through this process and then subject to hotspot analysis, and exploration of both spatial and temporal patterns of the topic discussion through the campaign. The data was also coupled with constituency estimates of the referendum vote to explore any relationships that may have been present between the two. The paper focuses on some of the key elements that arose from the study.

KEYWORDS: Brexit, Twitter, Topic Modelling, Spatial Analysis, R

1. Introduction

The UK referendum on membership of the EU generated a large amount of activity of social media, both in reaction and during the campaign period leading to the vote. As with all political campaigns, they are shaped by events, they drive discussion and whether scripted or random, events are captured on social media platforms. Exploration of events on social media has a vast resource of research, generally focusing on tweets, which can be used to generate analysis on both temporal and spatial scales in reaction to events (Shelton et al, 2014; Li et al, 2017).

The availability and ability to explore tweets over the campaign period raised the following questions relating to the referendum:

1) How did topics during the referendum campaign change and did this change spatially?
2) Did twitter highlight any events that appeared to shape the campaign?
3) Is it possible to infer any relationships between the intensity of twitter topic discussion and votes at a constituency level?

To answer these questions, geolocated tweets and referendum result data were required, combining them with a variety of analytical techniques to generate the required results, methods that will be further explained in the following section. The short nature of this paper limits the detail and the topics that could be explored, so has been restricted to a couple of key aspects from the study.

2. Methods

The author was allowed access to a collection of tweets that covered the UK from a large time period that included the referendum campaign. A period starting from Mid-February 2016 (when the referendum date was announced) and concluding at the end of June 2016 (following the referendum result) was selected, allowing the whole campaign and associated events to be captured, whilst also capturing reaction to the result in the week that followed. Tweets were selected by filtering for key words and hashtags relating to the referendum, these terms were decided upon by exploring previous studies into the referendum (Llewellyn and Cram, 2016; Hürlimann et al, 2016). A sample of random tweets of equivalent number was also selected for the same period, whose purpose was to serve as a
proxy for the twitter population. Both sets of tweets were then filtered further to remove duplicate tweets (which may have been generated through the iterative process of searching tweets by term) and those from probable spam accounts. The sample tweets were also restricted to one per user, again to make the those more reflective of overall twitter population. The selected tweets were then split into weeks, Friday to Thursday, enabling reaction following the result to be easily determined.

The objective of the topic modelling process was simply to classify tweets into topics that would contain tweets related to that topic. This utilised the R packages tm (Feinerer and Hornik, 2017) and topicmodels (Grünn and Hornik, 2017). A Latent Dirichlet Allocation (LDA) model was applied for extracting topics from tweets, with tweet text transformed to comply with set parameters, such as search terms and stopwords removed from all tweets. To determine the optimal number of topics per week, a sample of the tweets were subjected to iterative testing of the topic model before settling on a value for that week. This allowed for the number of topics to vary between weeks where activity could be expected to vary. Once the optimal value was calculated, the topic model was run for all selected tweets for that week, with the result being each tweet classified a value for the topic it belonged to. A wordcloud was then generated for the aggregated text of each topic, providing a context for manual naming of each topic value, an example of which can be observed in Figure 1. The manual naming process was felt necessary as using frequent words may have generated too many topics to analyse and may not have been a true reflection. However, it should be considered that the manual process also had issues with regards to time spent classifying assigned values, as well as the issues of subjectivity and bias that may be considered a factor. Topics captured through the modelling process but not considered relevant to the study were discarded, with events including Eurovision and the European Football Championships occurring during the study period.

Tweets for each topic per week were again limited to one per user, again with the aim of being more reflective of the twitter population and reduce the impact of dominant users.

Figure 1 Wordcloud for the topic classified as being about David Cameron.

Here, the locations of tweets were then used to create a point feature for each topic per week, which could then be aggregated to areas as required through a point in polygon process. Two area aggregations were chosen to satisfy the questions posed, GB constituencies to allow for comparison with referendum result estimates, and a hexagonal grid more representative than other shapes (Carr, Olsen and White, 1992; Shelton et al, 2014). For both aggregations, an odds ratio (OR) was generated utilising the sample tweets, which was used to gauge the relative activity of a topic for a given area. The constituency results were then combined with estimated referendum results by constituency (Hanretty, 2016), as results on the night were not always collected at a constituency level, so some modelling was required to calculate constituency results.

The two aggregations were then subjected to analysis using GeoDa, with spatial weights calculated using queens contiguity, enabling the the calculation of Moran’s I Values and the generation of Getis-
Ord G* maps to display hotspots of tweet activity for each topic for the gridded aggregation.

3. Results and Analysis

Overall, 291,797 tweets were classified as related to the referendum, into 111 separate topics made up of 102,941 unique twitter users. It is also important to be aware of the distribution of the twitter population, with Figure 2 highlighting that areas of high clustering of the sample users can be generally seen to located around major urban centres, unsurprising as younger users who represent a high proportion of users on the platform on the platform (Longley, Adnan and Lansley, 2015). As can be observed in Figure 3, the number of tweets classified varied over the campaign period, with the most active period of the study observed around the referendum date. The most common topics are shown in Table 1, with the debates undertaken as part of the campaign being the most common topic.

Figure 2 Getis-Ord G* Hotspot Map of Sample Tweets for study period (February – June 2017)
Figure 3 Number of tweets classified to a topic per day for the study period.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Number of Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Debate</td>
<td>13324</td>
</tr>
<tr>
<td>Trump</td>
<td>12946</td>
</tr>
<tr>
<td>Trade</td>
<td>10990</td>
</tr>
<tr>
<td>Cameron</td>
<td>10696</td>
</tr>
<tr>
<td>Rights</td>
<td>9877</td>
</tr>
<tr>
<td>Scottish Independence</td>
<td>9151</td>
</tr>
<tr>
<td>Elections</td>
<td>8578</td>
</tr>
<tr>
<td>Stay</td>
<td>7890</td>
</tr>
<tr>
<td>Drop in Pound</td>
<td>6395</td>
</tr>
<tr>
<td>NHS</td>
<td>6206</td>
</tr>
</tbody>
</table>

The debates were clearly key events of the campaign, whilst they were classified as a topic on their own merit, they can also be observed to driven the increase in discussion of other topics, as Figure 4 shows for the final TV debate of the campaign.
It is also interesting to consider how the discussion of debates evolved spatially over the final weeks of the campaign, with Figure 5 displaying how the discussion was initially limited to major centres in week 15, with 2381 cells containing no tweets classified as debate related, a 300% increase in empty cells compared with the sample tweets used as a proxy for twitter population. The evolution and increase in cells discussing the debates as we draw closer to the referendum date in week 18 is clear, with areas of high clustering (typically areas of high twitter population) being increasingly linked by areas not considered significant, but still containing tweets about the topic.

Another aspect that appears from the study, is the potential appearance of a switch in users following the referendum result. This is supported by Figure 6, were constituencies with an increased leave vote, displayed a decreased OR after the referendum result compared to the before the vote. For most topics following the referendum, the number of cells showing activity is higher than for topics occurring before the referendum. Another example of this is the topic of a second referendum that was a strong
topic in reaction to the result, with Figure 7 showing the extent of discussion, it compares well in terms of cells showing activity with that of the final debate week, which is suggestive of wider engagement with the overall twitter population, and that discussion before the referendum was restricted to a subsection of users on the platform.

Figure 6 Change in OR between tweets classified before the result and those after, against the estimated percentage vote for leave in the referendum.
4. Conclusion

It can be clearly observed that the debates drove discussion on the platform, and is just one of the topics extracted through the topic modelling process. These events can also be seen to evolve and vary spatially over time, with differing users engaging with discussion on the platform before and after the event. It can also be seen that it is possible to infer relationships between how a constituency voted and twitter topics, although this again requires more robust testing, perhaps through using a GWR approach. The use of topic modelling for extracting opinion perhaps requires more refinement, especially with regards to topic classification. Finally, it is important to remember that this study reflects twitter users, and as such has limitations of representativeness, and should be framed in this context in further analysis.
5. Acknowledgements

I would like to thank Dr Stefano De Sabbata giving me access to the twitter data required for this study, his feedback and support when I approached him with the idea for this project. I’d also like to thank Professor Christopher Hanretty for making his estimates of constituency level results from the referendum publicly available.

6. Biography

Robert Webster has recently completed his MSc in Geographical Information Science at the University of Leicester, and was until recently working for Sterling Geo.

References


